# STT 873 HW1 (Solution Keys)

This HW is due on Sep 18th.

**Ex. 2.2**: Show how to compute the Bayes decision boundary for the simulatoin example in Figure 2.5. *(10 pts)*

**Solution:** The Bayes classifier is

$$\hat{G}(X) = \underset{g \in \mathcal{G}}{\operatorname{argmax}}\, P(g|X = x)$$

In this two-class example ORANGE and BLUE, the decision boundary is the set where

$$P(g = \text{BLUE}|X = x) = P(g = \text{ORANGE}|X = x) = \frac{1}{2}$$

By the Bayes rule, this is equivalent to the set of points where

$$P(X = x|g = \text{BLUE})P(g = \text{BLUE}) = P(X = x|g = \text{ORANGE})P(g = \text{ORANGE})$$

and since we know $P(g)$ and $P(X = x|g)$, the decision boundary can be calculated explicitly.

**Ex. 2.7**: Suppose we have a sample of $N$ pairs $x_i$, $y_i$ drawn i.i.d. from the distribution characterized as follows:

$$\begin{aligned} x_i &\sim h(x), \text{ the design density} \\ y_i &= f(x_i) + \varepsilon_i, \ f \text{ is the regression function} \\ \varepsilon_i &\sim (0, \sigma^2) \text{ mean zero, variance } \sigma^2 \end{aligned}$$

We construct an estimator for $f$ *linear* in the $y_i$,

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i,$$

where the weights $l_i(x_0; \mathcal{X})$ do not depend on the $y_i$, but do depend on the entire training sequence of $x_i$, denoted here by $\mathcal{X}$.

(a) Show that linear regression and $k$-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $l_i(x_0; \mathcal{X})$ in each of these cases. *(3 pts)*

(b) Decompose the conditional mean-squared error

$$E_{\mathcal{Y}|\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

into a conditional squared bias and a conditional variance component. Like $\mathcal{X}, \mathcal{Y}$ represents the entire training sequence of $y_i$. *(2 pts)*

(c) Decompose the (unconditional) mean-squared error

$$E_{\mathcal{Y},\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

intor a squared bias and a variance component. *(2 pts)*

(d) Establish a relationship between the squared biases and variances in the above two cases. *(3 pts)*

**Solution:**

(a) Recall that the estimator for $f$ in the linear regression case is given by

$$\hat{f}(x_0) = x_0^T \hat{\beta}$$

where $\hat{\beta} = (X^T X)^{-1} X^T y$. Then we can simply write

$$\hat{f}(x_0) = \sum_{i=1}^{N} (x_0^T (X^T X)^{-1}) X^T)_i y_i.$$

Hence

$$l_i(x_0; \mathcal{X}) = (x_0^T (X^T X)^{-1} X^T)_i.$$

In the $k$-nearest-neighbor representation, we have

$$\hat{f}(x_0) = \sum_{i=1}^{N} \frac{y_i}{k} 1_{\{x_i \in N_k(x_0)\}}$$

where $N_k(x_0)$ represents the set of $k$-nearest-neighbors of $x_0$. Clearly,

$$l_i(x_0; \mathcal{X}) = \frac{1}{k} 1_{\{x_i \in N_k(x_0)\}}$$

(b)

$$\begin{aligned}
E_{\mathcal{Y}|\mathcal{X}}[(f(x_0) - \hat{f}(x_0))^2] &= E_{\mathcal{Y}|\mathcal{X}}[(f(x_0) - E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) - \hat{f}(x_0))^2] \\
&= E_{\mathcal{Y}|\mathcal{X}}[(f(x_0) - E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)))^2] + E_{\mathcal{Y}|\mathcal{X}}[(E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) - \hat{f}(x_0))^2] \\
&\quad +2E_{\mathcal{Y}|\mathcal{X}}[(f(x_0) - E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)))(E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) - \hat{f}(x_0))] \\
&= \mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Bias}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))^2
\end{aligned}$$

(c) Here we simplify the notation $E_{\mathcal{Y},\mathcal{X}}$ to $E$.

$$\begin{aligned}
E[(f(x_0) - \hat{f}(x_0))^2] &= E[(f(x_0) - E(\hat{f}(x_0)) + E(\hat{f}(x_0)) - \hat{f}(x_0))^2] \\
&= E[(f(x_0) - E(\hat{f}(x_0)))^2] + E[(E(\hat{f}(x_0)) - \hat{f}(x_0))^2] \\
&\quad +2E[(f(x_0) - E(\hat{f}(x_0)))(E(\hat{f}(x_0)) - \hat{f}(x_0))] \\
&= \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}(\hat{f}(x_0))^2
\end{aligned}$$

(d) In (b) we have

$$\begin{aligned}
E[(f(x_0) - \hat{f}(x_0))^2] &= E_{\mathcal{X}}(E_{\mathcal{Y}|\mathcal{X}}[(f(x_0) - \hat{f}(x_0))^2]) \\
&= E_{\mathcal{X}}(\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)) + \mathrm{Bias}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))^2) \\
&= E_{\mathcal{X}}(\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))) + E_{\mathcal{X}}(\mathrm{Bias}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))^2) \qquad (1)
\end{aligned}$$

and in (c) we have

$$E[(f(x_0) - \hat{f}(x_0))^2] = \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}(\hat{f}(x_0))^2 \qquad (2)$$

Comparing (1) and (2) we have

$$\begin{aligned}
E_{\mathcal{X}}(\mathrm{Bias}(\hat{f}(x_0))^2) - \mathrm{Bias}(\hat{f}(x_0))^2 &= \mathrm{Var}(\hat{f}(x_0)) - E_{\mathcal{X}}(\mathrm{Var}_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))) \\
&= \mathrm{Var}_{\mathcal{X}}(E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0))) \\
&\geq 0
\end{aligned}$$

The above inequality suggests that the expectation of conditional squared bias of $\hat{f}(x_0)$ is always greater than or equal to the squared bias of $\hat{f}(x_0)$, and the difference is equal to $\mathrm{Var}_{\mathcal{X}}(E_{\mathcal{Y}|\mathcal{X}}(\hat{f}(x_0)))$.

**Ex. 2.8**: Compare the classification performance of linear regression and $k$-nearest neighbor classification on the **zipcode** data. In particular, consider only the 2's and 3's, and $k = 1, 3, 5, 7$ and 15. Show both the training and test error for each choice. The **zipcode** data are available from the book website **www-stat.stanford.edu/ElemStatLearn**. *(10 pts)*

**Solution:** The implementation in R (see appendix) and graphs are attached. It's clear that for $k = 1, 3, 5, 7$ and 15, the $k$-nearest neighbor has a smaller classification error for the testing dataset compared to that of the linear regression. Also note that the $k$-nearest neighbor classification error increases with $k$ for both training and testing datasets.

| Model | Training error | Test error |
|---|---|---|
| Linear Reg | 0.0058 | 0.0412 |
| 1-NN | 0.0000 | 0.0247 |
| 3-NN | 0.0050 | 0.0302 |
| 5-NN | 0.0058 | 0.0302 |
| 7-NN | 0.0065 | 0.0330 |
| 15-NN | 0.0094 | 0.0385 |

**Ex. 2.9**: Consider a linear regression model with $p$ parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1)$,. . . ,$(\tilde{x}_M, \tilde{y}_M)$ drawn at

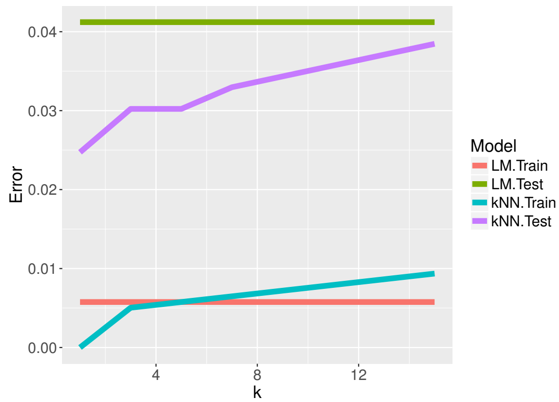


Figure 1: Classification errors for different methods on zipcode data.

random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \beta^T \tilde{x}_i)^2$, prove that

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})],$$

where the expectations are over all that is random in each expression. *(10 pts)*
**Solution:** Consider two cases:
(i) If $N \leq M$:

$$\begin{aligned}
E[R_{te}(\hat{\beta})] &= E\left(\frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2\right) \\
&= \frac{1}{M}\sum_{i=1}^{M} E(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2 \\
&\geq \frac{1}{M}\sum_{i=1}^{M} E(\tilde{y}_i - \tilde{\beta}^T\tilde{x}_i)^2 \qquad \text{where } \tilde{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{M}\sum_{i=1}^{M} E(\tilde{y}_i - \beta^T\tilde{x}_i)^2 \\
&= E(\tilde{y}_1 - \tilde{\beta}^T\tilde{x}_1)^2 \qquad \because (\tilde{x}_i, \tilde{y}_i)\text{'s are i.i.d} \\
&= \frac{1}{N}\sum_{i=1}^{N} E(\tilde{y}_i - \tilde{\beta}^T\tilde{x}_i)^2 \qquad \text{i.i.d again} \\
&\geq \frac{1}{N}\sum_{i=1}^{N} E(\tilde{y}_i - \tilde{\beta}'^T\tilde{x}_i)^2 \qquad \text{where } \tilde{\beta}' = \operatorname*{argmin}_{\beta} \frac{1}{N}\sum_{i=1}^{N} E(\tilde{y}_i - \beta^T\tilde{x}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} E(y_i - \hat{\beta}^T x_i)^2 \qquad \because (\tilde{x}_i, \tilde{y}_i)\text{'s and } (x_i, y_i)\text{'s are i.i.d} \\
&= E\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^T x_i)^2\right) \qquad \text{where } \hat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{N}\sum_{i=1}^{N} E(y_i - \beta^T x_i)^2 \\
&= E[R_{tr}(\hat{\beta})]
\end{aligned}$$

(ii) If $N > M$:

$$\begin{aligned}
E[R_{tr}(\hat{\beta})] &= E\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^T x_i)^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} E(y_i - \hat{\beta}^T x_i)^2 \\
&= E(y_1 - \hat{\beta}^T x_1)^2 \\
&= \frac{1}{M}\sum_{i=1}^{M} E(y_i - \hat{\beta}^T x_i)^2 \\
&\leq \frac{1}{M}\sum_{i=1}^{M} E(y_i - \hat{\beta}'^T x_i)^2 \qquad \text{where } \hat{\beta}' = \operatorname*{argmin}_{\beta} \frac{1}{M}\sum_{i=1}^{M} E(y_i - \beta^T x_i)^2 \\
&= \frac{1}{M}\sum_{i=1}^{M} E(\tilde{y}_i - \tilde{\beta}^T \tilde{x}_i)^2 \\
&\leq \frac{1}{M}\sum_{i=1}^{M} E(\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2 \\
&= E\left(\frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}^T \tilde{x}_i)^2\right) \\
&= E[R_{te}(\hat{\beta})]
\end{aligned}$$